Semantic annotation of French corpora: animacy and verb semantic classes

Juliette Thuilier, Laurence Danlos

Univ Paris Diderot, Sorbonne Paris Cité, ALPAGE, UMR-I 001 INRIA,
F-75013 Paris, France
jthuilier@linguist.jussieu.fr, laurence.danlos@linguist.jussieu.fr

Abstract

This paper presents a first corpus of French annotated for animacy and for verb semantic classes. The resource consists of 1,346 sentences extracted from three different corpora: the French Treebank (Abeillé and Barrier, 2004), the Est-Républicain corpus (CNRTL) and the ESTER corpus (ELRA). It is a set of parsed sentences, containing a verbal head subcategorizing two complements, with annotations on the verb and on both complements, in the TIGER XML format (Mengel and Lezius, 2000). The resource was manually annotated and manually corrected by three annotators. Animacy has been annotated following the categories of (Zaenen et al., 2004). Measures of inter-annotator agreement are good (Multi-$\pi = 0.82$ and Multi-$\kappa = 0.86$ (k = 3, N = 2360)). As for verb semantic classes, we used three of the five levels of classification of an existing dictionary: *Les Verbes du Français* (Dubois and Dubois-Charlier, 1997). For the higher level (generic classes), the measures of agreement are Multi-$\pi = 0.84$ and Multi-$\kappa = 0.87$ (k = 3, N = 1346). The inter-annotator agreements show that the annotated data are reliable for both animacy and verbal semantic classes.

Keywords: manual annotation, animacy, verb semantics

1. Introduction

In this paper, we present a first corpus of French annotated for animacy and for verb semantic classes. The resource consists of 394 sentences from the French Treebank (FTB, (Abeillé and Barrier, 2004)), 622 sentences from the corpus Est-Républicain corpus (ER)[1] and 330 from the ESTER corpus[2]. This resource was manually annotated and manually corrected.

2. The corpus and the annotation

The constitution of the corpus ties in with the research question we are interested in, namely the order of post-verbal complements in French. That order is relatively free as shown in example (1) and (2). Aside from the relative length of the constituents, we aim to find which constraints affect the choice between NP-PP or PP-NP order.

(1) il montrait [aux copains]PP [son butin]NP.

(2) il montrait [son butin]NP [aux copains]PP.
(from Est-Républicain Corpus)
"*he showed his loot to his friends*"

Studies dealing with constituent order in English have demonstrated that animacy and verb semantics play a role in dative alternation (Gries, 2003; Bresnan et al., 2007, among others) and that animacy also affects genitive alternation (Rosenbach, 2005).

In order to study the effect of these two constraints in French, we created a database of 1,346 sentences extracted from three corpora (FTB, ESTER, ER) that contains a ditransitive verb followed by only two complements subcategorized by the verb.

First, the sentences were automatically extracted from FTB using the already annotated syntactic structure and grammatical functions. Second, the sentences from ER and ESTER were manually selected according to the verbal head. Therefore, they were automatically parsed and the treebanks were manually corrected.

This database was annotated for the animacy of both complements' referents, following the categories of Zaenen et al. (2004) adapted to French. The semantics of the ditransitive verbs was annotated using an existing resource: "*Les Verbes du Français*" (LVF, (Dubois and Dubois-Charlier, 1997)). The Salto Annotation Tool (Burchardt et al., 2006) was used for the annotation process because it is an easy to use tool for graphical annotation of treebanks. Thus, the database is a set of parsed sentences with annotations on the verb and on both complements, in the TIGER XML format (Mengel and Lezius, 2000). Figure 1 shows an example of an annotated sentence visualized with the Salto tool.

3. Animacy

Animacy is an inherent semantic property of referents. Animacy is often conceived as a hierarchical property going from human to inanimate. In this work, the hierarchy that we used is presented in Table 1 and inspired by Garretson (2004).

HUMAN	>	ANIMATE	>	INANIMATE
human		animal		concrete
		organization		non-concrete
				machine
				vehicle
				place
				time

Table 1: Hierarchy of animacy (Garretson, 2004)

Even though much of the work based on animacy coded data uses simpler distinctions (e.g. animate vs. inanimate), such an elaborate hierarchy is useful for the anno-

[1] Freely available at `http://www.cnrtl.fr/corpus/estrepublicain/`. We used the lemmatized version of this corpus, which will be presented at LREC 2012 (Seddah et al., 2012) and which is freely available at `http://alpage.inria.fr/estrepu/` .

[2] Distributed by ELRA.

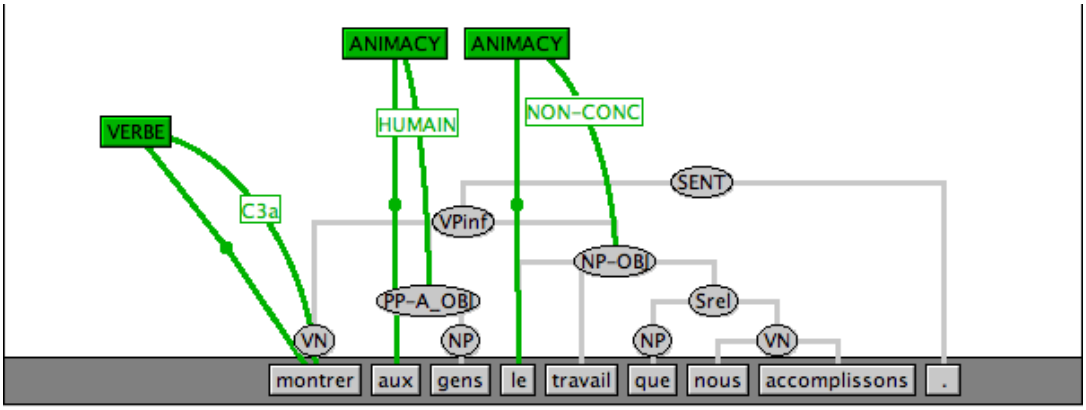

Figure 1: Example of annotated sentence visualized with Salto tool.

tation task. For instance, the distinction between *organization* and *non-concrete inanimate* referent is often difficult and uncertain. However, *organization* is categorized as animate, whereas *non-concrete* falls into the inanimate class. Thus, the discrimination between organization and non-concrete appears to be essential. We believe that using fine-grained categories with detailed definitions gives the annotator more elements to classify the difficult cases, and consequently, provides better annotation quality.

Three annotators carried out the annotation task: one of them was the author; the other two were undergraduate students in computational linguistics at Université Paris 7. They annotated the referent of both complements in the 1,346 sentences of the database using annotation guidelines (in French). Note that 332 complements are propositions introduced by a complementizer, so they are not relevant from the animacy point of view. The database for animacy is then composed of 2 360 referents (2 x 1014 + 332). First, the annotators worked independently. Then, during the adjudication process, they discussed the annotations on which they disagreed in order to reach consensus. So, the result of the annotation consists of three annotated corpora and one adjudicated corpus.

Following the terminology of Artstein and Poesio (2008), we report agreement in terms of Fleiss Multi-π (also known as Carletta's kappa (Carletta, 1996)) and Multi-κ (generalization of Cohen's (Cohen, 1960)), in order to estimate the reliability of the annotation with three annotators: Multi-$\pi = 0.82$ and Multi-$\kappa = 0.86$ (k = 3, N = 2360). The two measures indicate a good inter-annotator agreement, given that values higher than 0.8 are considered as reliable. In order to give an idea of the distribution of the categories, Table 2 contains the frequency with which each category is classified in the same way by the three annotators (the diagonal of the table) and the frequency of disagreement (other boxes). More precisely, considering the coders one pair at a time, Table 2 is a confusion matrix that displays the frequency with which one coder of the pair chose the category named in the row header while the other chose the category named in the column header for the same referent. We observe that the categories non-concrete, organization and human are the ones with most disagreement. These disagreements are similar to those observed by Zaenen et al. (2004) and can be understood as reflecting differences of interpretation of the constituent in the context. In the following subsections, we present two typical cases that annotators disagreed on.

3.1. Organization vs. Non-concrete

As mentioned earlier, the distinction between *organization* and *non-concrete* is essential, because *non-concrete* referents are categorized as *inanimate* while *organizations* are considered as *animate*.

(3) le gouvernement de M. Pierre Bérégovoy et M. Gomez [...] cèdent [l'usine Eisswein et l'électroménager de Thomson SA]NP [à un groupe familial étranger, l'italien Elettro Finanziaria Spa]PP (FTB)
litt: "*the government of Mr Pierre Bérégovoy and Mr Gomez sells the Eisswein factory and the electrical goods industry Thomson SA to the foreigner family group, the Italian Elettro Finanziaria Spa*"

(4) il fait [du groupe français]PP [le numéro un mondial en équipements de transmissions] (FTB)
litt: "*it makes of the French group the leader in engineering of transmission*"

In sentence (3) and (4), two annotators chose the tag *organization* for the PPs, considering that "*groupe familial*" and "*groupe français*" refer to a group of humans, whereas the other coder used the tag *non-concrete* considering that they refer to an abstract entity 'company'. In the adjudicated version of the corpus, we went for *organization* in sentence (3), since the beneficiary of the transaction seems to be the persons heading the company. As for sentence (4), we chose the *non-concrete* tag because we considered the referent to be the abstract business entity more than the organized group of persons.

3.2. Human vs. Organization

We observe disagreements between *organization* and *human*, because, in some contexts, it is difficult to say if the referent is more a human or a group of humans.

(5) la firme Trasgo fournit [des poussins]NP [à des ejidos chargés de les nourrir pendant huit semaines]PP

	Ani	Conc	Hum	Place	Non-conc	Orga	Time	Veh	Mach	Oanim
Animal	18	4	2	0	0	0	0	0	0	0
Concrete		181	0	21	122	3	0	4	6	0
Human			1767	2	112	108	0	0	0	8
Place				188	51	6	0	0	0	0
Non-conc					3807	214	62	0	8	23
Organization						251	0	0	0	5
Time							103	0	0	0
Vehicle								2	2	0
Machine									0	0
Oanim										0

Table 2: Confusion matrix for animacy (Oanim stands for 'I don't know')

(FTB)
litt: "*the Trasgo firm supplies chicks to ejidos responsible for feeding them for 8 weeks*"

In sentence (5), the PP can be interpreted as referring either to communities owing communal lands in Mexico or to the community members which actually feed the chicks. Two annotators used the *organization* tag and the other one the *human* tag. We chose *organization* for the adjudicated corpus, considering that the PP refers more to the community than to the members.

4. Verb semantic classes

The "*Les Verbes du Français*" dictionary (Dubois and Dubois-Charlier, 1997) is a hand-written resource containing 25,610 verbal entries, representing 12,310 verbs classified according to their syntactico-semantic properties. This dictionary is a very detailed resource, with a large coverage. The classification is based on the analysis of the types of subjects, complements and adjuncts (animate, inanimate, abstract, singular/plural, collective…), the realizations of the arguments (NP, PP, clause…), as well as the syntactic alternations allowed by the verb.
We used three of the five levels of classification for the annotation:

- 14 generic classes represented by an uppercase letter;
- 54 semantico-syntactic classes labeled with a digit;
- 248 syntactic sub-classes labeled with a lowercase letter.

The 14 generic classes indicate the general meaning of the verb. They are:

- **C**: communication
- **D**: donation/deprivation
- **E**: entrance/exit
- **F**: to hit/to touch
- **H**: physical condition/behavior
- **L**: locative
- **M**: movement in place
- **N**: to provide/to remove
- **P**: psychological verbs
- **R**: achievement/setting up
- **S**: to grab/to grip/to own
- **T**: transformation/change
- **U**: to combine/to bring together
- **X**: auxiliary verbs

The semantico-syntactic classes are generally arranged according the type of subject and the use of the verb (literal or figurative sense), except for C, D, P and X generic classes. This means that we need the generic class in order to interpret the semantico-syntactic level.

- E, F, H, L, M, N, R, S, T, U
  - **1**: human or animal, literal sense
  - **2**: human, figurative sense
  - **3**: inanimate, literal sense
  - **4**: inanimate, figurative sense
- D (donation/deprivation)
  - **1**: human
  - **2**: non-human, literal sense
  - **3**: non-human, figurative sense
- C (communication)
  - **1**: human or animal (to shout, to speak)
  - **2**: human (to say something)
  - **3**: human (to show)
  - **4**: figurative sense
- P (psychological verbs)
  - **1**: human subject
  - **2**: human object
  - **3**: inanimate object
- X (auxiliaries)
  - **1**: temporal or aspectual auxiliaries

	C	D	E	F	H	L	M	N	O	P	R	S	T	U	X
C	1435	56	0	2	0	4	0	0	0	2	2	0	0	2	0
D		760	16	6	2	10	0	2	0	15	2	37	2	4	0
E			241	1	48	1	8	0	5	0	0	0	0	3	0
F				6	0	0	0	4	1	0	0	0	0	0	0
H					139	14	0	0	1	1	4	10	0	0	0
L						264	3	2	0	14	37	30	0	1	2
M							66	0	0	0	1	0	0	2	4
N								14	0	0	0	0	0	2	0
O									0	0	0	3	0	0	0
P										58	0	3	0	5	0
R											246	0	106	0	22
S												36	0	3	0
T													147	0	0
U														115	0
X															6

Table 3: Confusion matrix for generic classes of verbs (Overb stands for 'I don't know')

- **2**: impersonals
- **3**: synonyms of 'to be'+ time or place
- **4**: to finish and to begin

The syntactic sub-classes indicate the syntactic construction the verb appears in. Given the size of this level, we cannot give a detailed overview of it. We rather present an example of an annotation with the verb *céder* 'to sell' in sentence (6).

(6) elle cède [à celui-ci]PP [3,5% de la SGAB et 19,6% de la ACESA]NP (FTB)
litt: "*it sells to this one 3.5% of SGAB and 19.6% of ACESA*"

The verb is annotated **D2a**, which means that it has:

- the generic class D (standing for 'donation/deprivation');
- the semantico-syntactic class D2 (standing for 'to give something to somebody/to get something from somebody');
- the syntactic sub-class D2a (standing for 'to supply somebody with something').

The same three coders realized this annotation task using the online version of the dictionary[3] as annotation guidelines. They annotated the 1,346 verbs of the database and enriched the annotation guidelines, listing and explaining the main difficulties and annotation choices. Like the animacy corpus, the adjudicated corpus is the result of a consensus between the three annotators for the cases they disagreed on.
The main difficulty lies in the fact that the hand-written resource has not been conceived for an annotation task. Thus, uses of verbs found in corpora do not systematically correspond to lexical entries. For example, the database contains occurrences of the verb *mettre* 'to put' employed with predicative PPs, as in *mettre en valeur* 'to emphasize'. However, the LVF has no entry corresponding to this kind of meaning. Moreover, the LVF differentiates between very close meanings of a verb, and it is sometimes difficult to identify these differences in contextualized examples.
Considering the syntactic sub-classes, Multi-π = 0.76 and Multi-κ = 0.78 (k = 3, N = 1346). Given the number of categories (248) and the nature of the resource used for the annotation, the agreement between the 3 annotators seems reasonable. The confusion matrix (Table 3) was conceived in the same way as the one concerning animacy. It only contains the generic classes. The corresponding Multi-π and Multi-κ are respectively 0.84 and 0.87.

5. First observations and results

These annotated data have been used in linguistics-oriented studies, dealing with the problem of verbal complement order.
First, animacy seems to not be relevant in French. In the sub-corpus composed of the sentences containing two phrasal complements, 10.2% of the NP and 37.0% of the PP are animate (= human, organization, animal). As shown in Thuilier et al. (2011), when the relative length of both complements, the verbal lemma, and the collocation effect between the verb and the PP are controlled, animacy does not show a significant effect on the relative order of post-verbal complements in our data. This result seems to be confirmed by an experiment based on a questionnaire where the length effect was neutralized (Thuilier et al., 2011).
Second, we observe that the verbs associated with their semantic classes have different behavior according to verbal complements order. Thuilier (Forthcoming) points out that we can have a better modeling of the order of verbal complements when taking into account the disambiguated verb.

The measures of agreement indicate that the corpus presented here is a reliable semantically annotated resource for animacy and general semantic classes of verbs. It is a relevant data set from a linguistic point of view, as shown in the last section. Additionally, it can be used as a training set for automatic classification of semantic layers on treebanks.

[3] http://rali.iro.umontreal.ca/Dubois/

6. Acknowledgements

Alpage (INRIA Paris-Rocquencourt & Université Paris 7) supported this work by paying two coders for the annotation task. Thank you to these annotators, Kévin Deturck and Fabien Andreani, for their work.

7. References

Anne Abeillé and Nicolas Barrier. 2004. Enriching a french treebank. In *Proceedings of Language Ressources and Evaluation Conference (*LREC*)*, Lisbon.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistic*, 34:555–596, December.

Joan Bresnan, Anna Cueni, Tatiana Nikitina, and Harald Baayen. 2007. Predicting the dative alternation. In G. Boume, I. Kraemer, and J. Zwarts, editors, *Cognitive Foundations of Interpretation*. Royal Netherlands Academy of Science, Amsterdam.

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, and Sebastian Padó. 2006. SALTO – A Versatile Multi-Level Annotation Tool. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genoa, Italy.

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistic*, 22(2):249–254.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Jean Dubois and Françoise Dubois-Charlier. 1997. *Les verbes français*. Larousse-Bordas, Paris.

Gregory Garretson. 2004. Coding practices used in the project optimal typology of determiner phrases. http://npcorpus.bu.edu/html/documentation.

Stefan Th. Gries. 2003. Towards a corpus-based identification of prototypical instances of con- structions. *Annual Review of Cognitive Linguistics*, 1:1–27.

Andreas Mengel and Wolfgang Lezius. 2000. An xml-based representation format for syntactically annotated corpora. In *In Proceedings of the International Conference on Language Resources and Evaluation*, pages 121–126.

Anette Rosenbach. 2005. Animacy versus weight as determinants of grammatical variation in english. *Language*, 81(3):613–644.

Djamé Seddah, Marie Candito, Benoît Crabbé, and Enrique Henestroza Anguiano. 2012. Ubiquitous usage of a french large corpus: Processing the est republicain corpus. In *Proceedings of Language Ressources and Evaluation Conference (*LREC*)*, Istanbul.

Juliette Thuilier, Anne Abeillé, and Benoît Crabbé. 2011. Do animate arguments come first? In *Proceedings of Architectures and Mechanisms for Language Process (AMLAP 2011)*, Paris.

Juliette Thuilier. Forthcoming. Lemme verbal et classe sémantique dans l'ordonnancement des compléments postverbaux. In *Actes de CMLF 2010 - 2ème Congrès Mondial de Linguistique Française*, Lyon.

Annie Zaenen, Jean Carletta, Gregory Garretson, Joan Bresnan, Andrew Koontz-Garboden, Tatiana Nikitina, M. Catherine O'Connor, and Tom Wasow. 2004. Animacy encoding in english: why and how. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, DiscAnnotation '04, pages 118–125, Stroudsburg, PA, USA. Association for Computational Linguistics.